# PaLi vs CLIP: A Comparison of SOTA Visual-Language Multimodal Models

**Howard Tangkulung**[1]
[1]Faculty of Electrical and Electronics Engineering, The University of Tokyo, Japan

## Abstract

Mulitmodal tasks such as VQA remain a challenging task due to its high complexity and computational cost. In June 2022, Deuser, F., et al. (2022) [3] proposed a simple architecture of utilising CLIP models, reporting a VizWiz accuracy [1] of 0.6015. In September 2022, Chen, X., et al. (2023) [2] proposed the PaLI model which outperformed CLIP with a VizWiz accuracy of 0.7330. In this report, the proposed simple CLIP architecture and fine-tuned PaLI model are implemented and compared. When trained and tested on the VizWiz dataset provided, the PaLI model outperforms the CLIP model with a VizWiz accuracy of 0.72532 compared to 0.62516.

## 1 Introduction

### 1.1 VizWiz VQA Challenge

Visual Question Answering (VQA) is an artificial intelligence task that requires a model to answer visual questions. In particular, the VizWiz dataset comes from visually impaired users who took an image, recorded a question, and 10 crowdsourced workers answered the question. This challenge, known as the VizWiz VQA challenge was pulished in 2010, and since then has been a benchmark for visual-language multimodal models. As there are 10 answers per question, the VizWiz accuracy is defined as follows.

$$\text{VizWiz Accuracy} = min(\frac{\text{number of humans that provided that answer}}{3}, 1) \tag{1}$$

Hence, to achieve a VizWiz accuracy of 1, the model must provide the same answer as at least 3 humans. All accuracy scores in this paper are reported in VizWiz accuracy.

### 1.2 Models Overview

In this paper, 3 models are compared: the baseline model, CLIP (Contrastive Language-Image Pre-training) [6], and PaLI (Pre-trained and Attention-based Language-Image) [2]. For the baseline model, a simple architecture with ResNet-50 for image features and2 linear layers for text features is used. For the CLIP model, following the architecture proposed by Deuser, F., et al. (2022) [3], the CLIP model is used with 2 linear layers. Lastly, for the PaLI model, the 3B-PaliGemma model available in Hugging Face [4] is fine-tuned.

Final Assignment VQA Deep Learning Fundamentals 2024
Organised by Matsuo - Iwasawa Lab, Tokyo, Japan.

## 2 Model Architecture

### 2.1 Baseline Model

The baseline model, shown in Fig. 1,consists of a ResNet-50 model for image features and a linear layer for one-hot encoded text features.

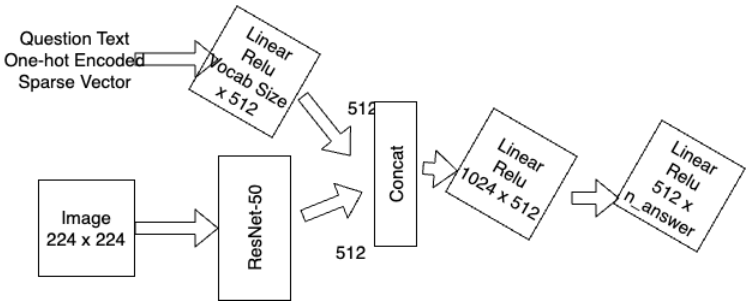



Figure 1: Baseline model architecture

### 2.2 CLIP Architecture

The CLIP model architecture, shown in Fig. 2, consists of a CLIP model with 2 linear layers. This model is based on the architecture proposed by Deuser, F., et al. (2022) [3], where

> A simple linear classifier is used on the concatenated features of the image and text encoder.

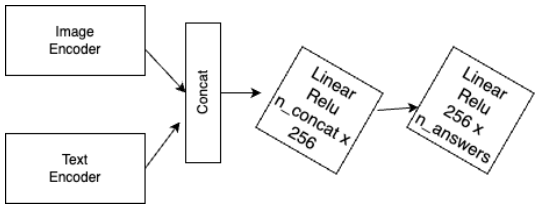



Figure 2: CLIP model architecture

In the proposed model, after linear layers, there is a AUX layer trained on answer mask. However, this layer is not used in this implementation as it is not available in the dataset provided. For this implementation, the CLIP model used is ViT-L/14@336px [5].

### 2.3 PaLI Architecture

The PaLI model architecture, shown in Fig. 3, consists of a 3B-PaliGemma model available in Hugging Face [4]. Instead of using linear layers to classify the image and text features, the PaLI architecture uses a decoder to generate the answer.

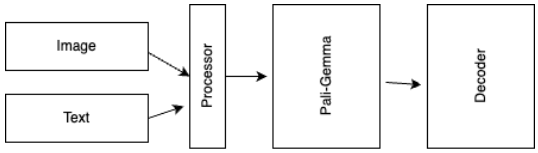


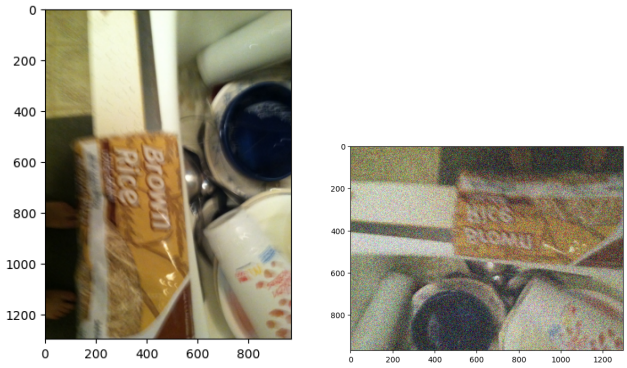

Figure 3: PaLI model architecture

# 3 Model Training

## 3.1 Preprocessing

### 3.1.1 Image Preprocessing

Image are resized to match the model input size. Image preprocessing methods such as data augmentation have proven uneffective in improving the model performance. This is due to the fact that the VizWiz dataset has low quality data and that important information may be lost during data augmentation. For example, consider the image in Fig. 4.




(a) (b)

Figure 4: Example of images in the VizWiz dataset

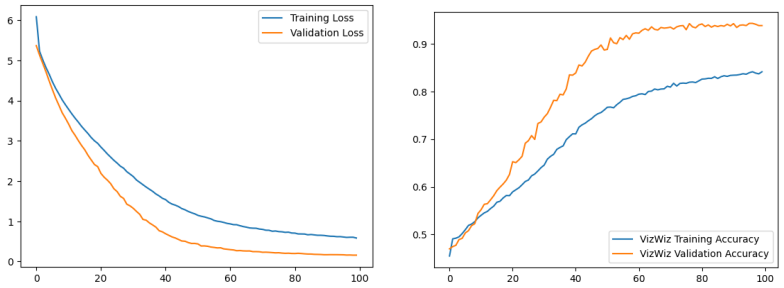
When asked the question "What is this?", the ideal answer will be "Brown Rice" but if the image is augmented, due to noise, the model might not be able to capture the important information (i.e. the packaging name). In this implementation, when the image is augmented in the CLIP architecture, the model performance decreased from 0.62516 to 0.57955. Hence, image augmentation does not improve the model performance.

### 3.1.2 Text Preprocessing

For each question, there are 10 answers provided by humans. To find the "best" answer for each question to be used in training, the most common answer is selected. However, if there is no majority answer (multiple answers with the same frequency), the answer with the least pair wise Levenshtein distance to other answers is selected. This process of finding the most representative answer is inspired by the work of F. Deuser et al. (2022) [3].

## 3.2 CLIP Model

The CLIP model is first imported from github [5]. Then, 2 linear layers is added to the model. For the final layer, a linear layer with output units the size of the number of unique answers in the dataset is added. Hence, the model will not predict answers that are not in the training dataset. Due to this limitation, in this implementation, a part of the training dataset is used as the validation dataset instead of using a separate validation dataset. The model is then trained on the VizWiz dataset provided, with CrossEntropyLoss as the loss function, Adam as the optimizer and a learning rate of 0.005, trained for 100 epochs.




(a) (b)

Figure 5: CLIP model loss and VizWiz accuracy

The loss and VizWiz accuracy is shown in Fig. 5. As shown, the model converges after 50 epochs.

### 3.3 PaLI Model

PaLI model, google/paligemma-3b-pt-224, is imported from Hugging Face [4]. To fine-tune the model, LoraConfig from PEFT is used to select target modules, such as Query Projection, Key Projection, Value Projection, and etc. Then, the model is trained using the lightning module, with a learning rate of 0.0001, trained for 10 epochs. The progression of VizWiz accuracy is shown in Fig. 6.

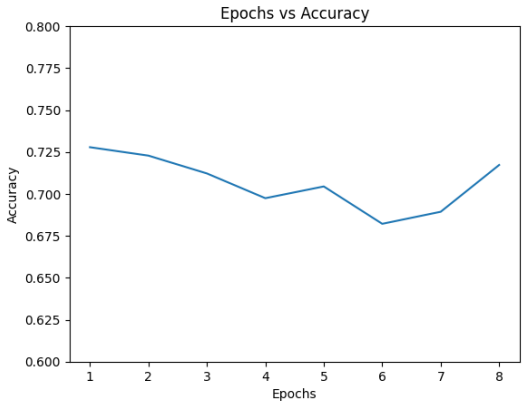



Figure 6: PaLI model test VizWiz accuracy

As shown, the model performs very well even with only 1-2 epochs of training. This is due to the fact that the model is already pre-trained on a large dataset and only needs to fine-tune on the VizWiz dataset. Additionally, the learning rate of 0.0001 could be too high for the model, as the model starts to overfit after 2 epochs.

## 4 Results and Discussion

For this assignment, two models are implemented and compared: the CLIP model, with a VizWiz accuracy of 0.62516, and the PaLI model, with a VizWiz accuracy of 0.72532. Implementation of the CLIP model followed the architecture proposed by Deuser, F., et al. (2022) [3], while the PaLI model is fine-tuned using the 3B-PaliGemma model available in Hugging Face [4]. CLIP models are known to be effective in visual-language multimodal tasks because it is pre-trained on unfiltered and highly noisy data. The training task for CLIP is given an image, the model has to predict which text (of a set) is the most likely pair. Hence, CLIP models are suitable for classification tasks. PaliGemma, on the other hand, takes image and a text string, such as a prompt, and generates a text string as output. This makes PaliGemma suitable for annotation generation tasks, such as VQA.

As the goal of this assignment is to score a high VizWiz accuracy, implementations are kept simple and only done on reported architectures. For a more comprehensive comparison, CLIP models should also be fine-tuned on its transformer layers. Additionally, the PaLI model could be further fine-tuned, by adjusting hyperparameters such as learning rate, batch size, and number of epochs.

## References


[1] VizWiz VQA Challenge. Visual question answering – vizwiz. https://vizwiz.org/tasks-and-datasets/vqa/. (Accessed on 06/22/2024).

[2] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. Pali-3 vision language models: Smaller, faster, stronger. https://arxiv.org/abs/2310.09199, 2023.

[3] Fabian Deuser, Konrad Habel, Philipp J. Rosch, and Norbert Oswald. Less is more: Linear layers on clip features as powerful vizwiz model. https://arxiv.org/abs/2206.05281, 2022.

[4] Google. Paligemma model card. https://huggingface.co/google/paligemma-3b-pt-224. (Accessed on 07/13/2024).

[5] OpenAI. Clip model card. https://github.com/openai/CLIP/tree/main. (Accessed on 07/13/2024).

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. https://arxiv.org/abs/2103.00020, 2021.